DOCUMENT RESUME

ED 279 681 TM 870 068

AUTHOR Hathaway, Walter E.
TITLE Toward an Ideal System of National Assessment of Educational Progress: A Local Perspective.
PUB DATE Aug 86
NOTE 55p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Testing; Elementary Secondary Education; *Item Banks; *Measurement Objectives; *National Surveys; *School Districts; State Surveys; Test Construction; *Testing Programs
IDENTIFIERS *National Assessment of Educational Progress; Portland School District OR

ABSTRACT

An ideal system of the National Assessment of Educational Progress (NAEP), from the local school district perspective, must follow several principles based on the 1985 Standards for Educational and Psychological Testing: (1) Testing must be viewed by teachers and students as worthwhile. (2) Test results must be presented in a timely and useful fashion. (3) Testing at national, state, and local levels should not be redundant. (4) Measurement should not disrupt normal classroom activities and attendance. (5) Reports should focus on factors over which the instructional leaders have some true control. (6) Tests should measure only curriculum outcomes which are being taught or which, it is agreed, should be taught. The ideal system would link national, state, and commercially available item banks. For each of the basic skill areas, five components would be included: (1) national learning objectives, (2) national public domain item banks, (3) a linking measurement scale, (4) a test development system, and (5) a testing or data collection system. The appendices, which constitute the greater part of the document, discuss federal government-school district relationships, data collection and decision-making plans, Oregon's school district profile, and an extensive presentation of the Portland (Oregon) School District's development of calibrated item banks ("A Regional and Local Item Response Theory Based Test Item Bank System," by Hathaway, Houser, and Kingsbury). (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279681

Toward an Ideal System of National Assessment of Educational Progress
A Local Perspective

Walter E. Hathaway

Director of Research
Portland Public Schools
Oregon

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. E. Hathaway

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

TM 810 068

BEST COPY AVAILABLE

Toward an Ideal System of National Assessment of Educational Progress

-

A Local Perspective

Dr. Walter E. Hathaway
Director of Research and Evaluation
Portland (OR) Public Schools

August, 1986

INTRODUCTORY

It is good and welcomed that local district perspectives are being sought as we work toward the redesign of the National Assessment of Educational Progress system. The current NAEP system has some severe theoretical and practical problems that have earned widespread criticism from local district staff whose cooperation is essential to its success. Moreover, discussions and planning are underway for possibly moving NAEP to the center of a more comprehensive national data collection system intended to serve state and local decision making purposes as well. (See Appendix I). Such plans to be successful require the enthusiastic and extensive participation of those who control access to the students whose progress is to be assessed and who make the most immediate and important decisions affecting students' academic performance.

An ideal system of National Assessment of Educational Progress from the local perspective must follow certain general principles of effective and efficient educational measurement and data collection that are of importance from our point of view. In this paper, I will attempt to lay out some of those principles and then I will sketch out an integrated, national-item-bank-based system that meets these principles. I will not at this time deal with other important but more particular issues such as:

- what to test
- when to test
- whom to test and whom to exclude
- what reference scales and reporting formats and procedures to use, etc.

PRINCIPLES OF EFFECTIVE AND EFFICIENT MEASUREMENT FROM A LOCAL PERSPECTIVE

In presenting these principles, I assume the application of appropriate standards from the Standards for Educational and Psychological Testing, 1985.

1. Any testing must be viewed by teachers and students as important and worth the effort to respond actively and appropriately.

2. Any test data gathered must be returned to the teachers and students involved in a timely fashion and the results must be viewed by them, as useful.

3. As far as possible, we should avoid redundant measurement for national, state, district, building, classroom, and student decision making purposes. (All these purposes need to be spelled out in advance and the measurement needs to be designed to fulfill them).

4. Measurement should not disrupt instruction and normal building and classroom activities any more than necessary. For example, whenever possible classrooms rather than single "pulled-out" students should be the sampling unit in sample-based testing and test results reporting.

5. Reports should focus on factors over which the instructional leaders have some real control. The most critical accountability factor for the instructional leaders is how effectively they can adapt the learning environment to meet the needs and abilities of individual students. A measurement program supporting longitudinal data can provide this type of data, but a program which only uses a single data point is quite limited. (See Appendix II)

6. The tests should measure only curriculum and learning outcomes that we are teaching or that we agree we should be teaching. Moreover, since the national tests will tend to establish a national curriculum and to focus the instructional efforts and time of those concerned with their relative standings on inevitable but undesirable state and local comparisons, it is essential that the outcomes covered by the test add up to as comprehensive as possible a definition of the desired measurable learning outcomes in an area at a level. Finally, efforts need to be made to educate the users of the test results on the curricular and other limitations of the tests.

Some local districts already pressed by state accountability developments view the development of a comprehensive national educational measurement system and its potential for a national curriculum as one more threat to "local control." Such districts would tend not to support the proposal that follows.

TOWARD AN IDEAL NAEP SYSTEM

From my local district perspective, an ideal National Assessment of Educational Progress would not be a test or a set of tests but an integrated system of advanced but proven psychometric and electronic technology. Such a system would:

- Be flexible enough to meet the needs for assessment information at all the following levels in an integrated and non-redundant fashion

  Local District
  student, parent

  teacher, instructional support staff

  principal, building management team (See Appendix III)

  central administrator, policy maker, community

  State System
  administrators, policy makers, and legislators

  National
  administrators, policy makers, and legislators

- Have as its primary function supporting teachers, students, parents, and support staff as they make placement, grouping, program, career, counseling, and instructional decisions that directly affect the individual learner. It would also stimulate and focus follow-up diagnosis and prescription to identify and meet individual student learning needs.

FIGURE I

State or local district developed tests and testing systems demonstrated to have an adequate match to a National consensus core curriculum, but possibly extending beyond it.
LINKED ←→
National Item Banks and Scales.
Covering National consensus core curriculum in each subject as well as additional theoretically or practically desirable outcomes.
From the National Item Banks.
National Tests can be developed for administration to representative samples of classrooms nationwide.
LINKED ←→
Major commercially available tests and testing systems demonstrated to have an adequate match to a National consensus core curriculum and and to the possibly broader curricula of the state and/or local districts being assessed.
Administered to a representative sample of classrooms at selected grades for selected subjects statewide (or to all classrooms).
Administered to all classrooms at selected grades for selected subjects in local districts.

- Be based upon a curricular consensus of the important measurable learning outcomes in each subject area, but with room for selection of additional outcomes beyond a national "core" to be measured at the state and local district levels. This core should be relatively narrowly defined and should include those instructional goals commonly taught in schools and found in the major published instructional materials. The item banks, however, should be quite comprehensive in their curriculum coverage so as to allow measurement even in the most uniquely designed curricula. These items should not, however, measure particular teaching strategies.

- Avoid over simplistic, destructive, politically motivated comparisons among state and for local school systems. (See Appendix IV).

In this system, students would be measured at selected grades in selected subjects for purposes of national decision making. For example, Reading, Language Usage, and Mathematics might be measured annually in grades 3 through 10; and Science, Writing Composition, and Social Studies might be measured at grades 5, 8, 10, and 12 in alternate years, etc. State and local school districts could adopt the national tests for their classroom decision making purposes. Alternately they could draw upon the national item banks to develop measures beyond the national core curriculum or could use linked commercially available instruments.

This ideal NAEP testing system would have five components in each of the basic skill areas:

1) a national set of learning objectives which would include a goal structure derived from a review of a representative sample of school districts nationwide. The learning objective should be more inclusive than the curriculum for any specific school district.

2) a set of national "public domain" item banks in the basic skills which would include items under each of the learning objectives defined in the national learning objectives. This item bank should consist of many items under each learning objective, in varied formats and item types. These items could be obtained from school districts, test publishers, and textbook publishers. Other items should be written to provide coverage of all learning objectives at all possible levels of difficulty. Each item bank, in whole or in

part, should be made available to local school districts, test publishers, etc. The existence of such a national item bank would foster the use of good measurement at all levels of the teaching and testing community.

3) a national measurement scale which would link all of items in the item banks. The existence of such a national measurement scale, developed using an appropriate item response model, would assure comparability of test scores from any test based on items from the national item bank. This scale would also allow for the development of local, state, and national norms, as desired. (Since multiple tests would be directly comparable, the collection of test data from multiple groups would be facilitated.)

4) a test development system which would be made available to test publishers, school districts, etc. This system would allow test developers to create tests from the national item banks which would be tailored to the specific measurement needs of the situation. This system should be easy enough for a novice to use, but powerful enough to allow the development of meaningful tests for varied purposes (including diagnostic testing, mastery testing, etc.).

5) a test delivery/data collection system which would allow NAEP to administer special purpose tests as necessary and collect data from users of the item banks. As a condition of the use of the item banks, NAEP should establish guidelines for the return of test data to enable the creation and update of item statistics and norming statistics. This would allow NAEP to compile nationwide statistics as desired. In order to facilitate the test delivery/data collection effort, an electronic version of the item banks should be created, with appropriate software for test printing and student data recovery (i.e., scanning and hard entry software). Data could be returned to NAEP by floppy disk or tape.

The key to making NAEP useful is to make the data useful to school districts while making the data collection as painless as possible. The development of a public domain item bank with appropriate software and data collection procedures would allow NAEP to perform immediately useful services to local districts while collecting data which would be useful for making state and national decisions.

In the development of such a national measurement system, we should minimize the emphasis on "national norms." Obtaining and maintaining true national norms is difficult and costly. Norms also tend to focus the measurement on a single point in time rather than on longitudinal trends which provide the most appropriate information for educational decision making. NAEP should focus on the development of "national scales" rather than "national norms."

We recognize the pressure for "national norms." The "ideal" system would not expend the tremendous resources necessary for the development of national norms when scaler data on the "core" may provide the information necessary for national decision making at far less cost. Critical points on the scales can be identified for decision making purposes through studying various decision requirements. Including the "core" items when states or local districts do survey measurement and then passing that data back to NAEP for sealing, may be sufficient for producing data for national decision making.

Such a system of measurement would improve both the quality and efficiency of educational measurement at reasonable cost. It would alleviate the dangerous emerging situation in which teachers are forced to test their students over and over for different purposes, e.g., once for themselves, and again for the local school board and administrator, once more the the state, and yet again to meet federal requirements.

The key alternative to an integrated, item-bank-based system such as the one proposed would be a well designed and sympathetically operated system of national measurement that would be as parsimonious and unburdensome as possible. Such a system should be based upon a minimal sample and should observe the general principles stated above.

REPERENCES

Ambach, Gordon M., Sheldon H. White, Richard J. Light. Harvard Educational Review, Vol. 52, No. 4, 1982.

"Analysis and Recommendations on Student Achievement Measures, "excerpted from CCSSO Center on Assessment and Evaluation, Draft Report of CEIR Committee, October 17, 1985.

Cooke, Charles, Alan Ginsburg, Marshall Smith. "The Sorry State of Education Statistics." Basic Education, Volume 29, No. 5, pp. 3-8., January 1985.

"Cross-National Education Indicators System (draft)." USDE, Office for Educational Research and Improvement Center for Statistics, Education Outcomes Division, Cross Sectional and Special Studies Branch. Contact Person: Eugene H. Owen.

David, Jane L. "Improving the Quality and Utility of NCES Data."

Hall, George, Richard M. Jaeger, C. Philip Kearney, David E. Wiley, "Alternatives for a National Data System on Elementary and Secondary Education." Prepared for The Office of Educational Research and Improvement, USDE, December 20, 1985.

"Plan for the Redesign of the Elementary and Secondary Data Collection Program (working paper)." Center for Statistics, Office of Educational Research and Improvement, USDE, March 27, 1986.

Plisko, Valena White, Alan Ginsburg, Stephen Chaikind. "Assessing National Data on Education." USDE, August 1985.

"The Sounding Board on Elementary/Secondary Statistics Redesign: Reactions to the Plan for the Redesign of the Elementary and Secondary Data Collection Program." Presented to The Committee for Coordinating Education Information and Research of The Council of Chief State School Officers by The Sounding on Elementary/Secondary Statistics Redesign convened by The Council of Chief State School Officers, June 1986.

Suter, Larry E. "Cross National Statistical Indicators of Education: Plans, Problems, and Possibilities." Presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 1986.

Walberg, Herbert J. "National Statistics to Improve Educational Productivity." University of Illinois at Chicago.

APPENDIX I

A Review of the March 27, 1986 Working Paper of the U.S. Department of Education "Plan for The Redesign of The Elementary and Secondary Data Collection Program" (The Plan)

Remarks Prepared for the Sounding Board of the CCSSO State Education Assessment Center on Elementary/Secondary Statistics Redesign

Walter E. Hathaway
Director of Research and Evaluation
Portland Public Schools (OR)
May 15, 1986

Many local school districts have carefully planned and developed student, school, and district information systems. These systems have been evolved over the last decade and more to efficiently respond to the information needs of our teachers, students, parents, support staff, administrators policy makers, and community. They include data on many of the indicators proposed both by the Council of Chief State School Officers (CCSSO) and by the U.S. Department of Education. Embedded in them are multi-layered systems of student performance assessment carefully aligned with the curriculum our teachers are teaching and designed to help them teach it to all their students. These information systems have earned the respect and trust of the constituencies they serve. They are at least partially responsible for the turnaround in the level of student performance in America which began in the mid-1970's and which today has brought our students' achievement to the highest levels in three decades, as documented in a recent Congressional Budget Office Report, Trends in Educational Achievement.

We, therefore, have a "show me" attitude toward state and federal proposals to replace or add additional layers on top of these effective local information systems with others whose chief virtue is uniformity permitting cross system comparisons. The cost of such replacement or additions is enormous in terms of the adjustments needed within even samples our over 15,000 school districts and the information and instructional time lost to our over 2.5 million teachers and other instructional staff and the 45 million students they serve. Moreover, the potential for misuse of poorly planned data is great.

We are, therefore, saying:

- Show us that you have some legitimate decision making needs for the cross system comparisons you wish to make that are proportionate to the cost to us and to you. (We have been made leery by the harm and demoralization caused by inappropriate judgments about being second rate such as were and are being made on the basis of the infamous and vastly over-simplified wall charts.)

- Show us that you know how to make such comparisons in a way that is valid and helpful given the incredible diversity among state and local school systems and the schools within them.

- Show us that you have the commitment, competence, patience, and resources to take the necessary time to do the job well and to design and operate a national educational data collection system that meets your own criteria of comprehensiveness, integration, accuracy, comparability, representativeness, timeliness, and minimal respondent burden in the judgment of a representative panel of qualified peer reviewers.

- Show us that you are not once again designing the means to impose on us the administration's ideological social and educational agenda which favors private over public education and promotes a one-sided view of "choice, content, and character."

- Finally and most importantly, show us that the system you propose will help improve teaching and learning i.e., that it will positively affect the content and teaching knowledge and skills of our teachers and their willingness and ability to consistently work hard with earned satisfactions and rewards to teach an appropriate curriculum; and that it will help students to consistently put out the maximum effort possible to benefit from such excellent teaching.

Because of the above concerns, I urge that the Center for Statistics consider the following:

1. Spell out the questions to be answered and the decisions to be informed by a national data system as well as any theoretical or conceptual frameworks employed.

2. Spell out the variables, data elements, and relations to be addressed and the criteria they must meet.

3. Solicit extensive and representative state and local input on 1 and 2.

4. Develop a detailed implementation plan.

5. Make the classroom the sampling unit rather than pull 84,000 students out of class repeatedly.

6. In general, proceed slowly and cautiously and with continuous peer (state and local) review at every step of the way. Don't operate in "political time: but in real time." Use some of this time to educate yourselves about the education information systems already in place or emerging in local school systems. Any useful national or state information systems must be designed to be compatible with and supportive of these local systems.

7. Don't by-pass the states but make them integral partners from the beginning and ensure that legitimate state level as well as national level policy issues are able to be appropriately informed.

8. Scale down your claims and expectations especially in the area of relational questions. Even a good national data system cannot answer <u>all</u> the important questions for the operation and improvement of education and indeed can make only a modest contribution to understanding and bettering learning.

9. Be aware of current theoretical and technical developments that promise to result in multi-level data that meets all your criteria. Emerging information systems operate through distributive processing networks of computers systems at the classroom and school levels. They permit the aggregation upward of real and accurate data for decision making at other levels such as district, state, and national.

PRINCIPLES IN REPORTING ACHIEVEMENT SCORES

APPENDIX II

---

1. Primary function of our reports is to support instructional decision making.

2. Data should provide information necessary to answer the question(s) being asked. (Reports should be designed around the questions.)

3. Provision of data only to staff members who have the instructional or administrative responsibility over the students or groups being summarized.

4. Avoid any possible evaluative focus on individual teacher performance.

5. Avoid encouragement for lowering fall scores or raising spring scores by artificial means so as to create the spurious appearance of growth.

6. Focus of attention on growth of students and groups.

7. Provision of longitudinal reference wherever possible.

8. Aim for simplicity but provide sufficient data for the decisions being made by each audience.

9. Graphic presentation whenever possible.

10. Focus on error band reporting.

11. Reports should encourage understanding of norms.

12. Reporting measures of central tendency and measures of variability reported together with the N used in computing them.

13. Layer the reporting package with optional reports of greater complexity for those who desire them.

14. Ranking of data in general may lead to misinterpretation.

15. Interpretation guides should be developed and distributed with each report.

16. When measuring educational achievement goals, the curriculum definitions must be the outline for the test blueprint. (The curriculum drives the measurement not the other way around.)

17. Graphic profiles are desirable in summarizing interactive relationships.

18. The overall size of the report and the clarity of the title should be made to be easily incorporated into current filing systems.

19. The accuracy and completeness of a report must be considered together with the timeliness and nature of the decisions being made from the data supplied in the report.

APPENDIX III

DECISION MAKING STRUCTURE

DECISION TYPE	TIME	TERM	REPORT TYPE	OTHER DATA	WHO	NEEDS/CONCERNS
1-- Instruc. Forecasting	Spring	MT		Port. Prognostic Markham	Principal	
-- Program Assessment	Spring	MT	GGP		Principal	
-- Placement	Fall	MT	Prelim. Princ.		Prin/Tcher	
-- Instruc. Planning	Fall	LT		BRAG/CC/CRT	Teacher	
-- Instruc. Monitoring	F-S	ST		GRAG/CAT/CRT	Teacher	
2-- Instruc. Emphasis	Spring	LT	Goal Report		Principal	
-- Instruc. Organiz.	Spring	LT	Goal Report		Principal	
-- Target Instruction	Spring	LT	Goal Report		Principal	
-- Expecta. for groups	Spring	LT		Class profile, CMI/grades teacher judgment	Teacher	
-- Instruc. Goals	May/Jun	LT		BRAG/ CRT Curriculum guide	Teacher	
-- Instruc. Groups	June	LT	Rank by grade		Prin/Tcher	
-- Reassign Groups	June	LT	Rank by class		Teacher	
-- Monthly Goals	June	LT	Goal Reports		Teacher	
-- Placement	Fall	MT	PALT	CAT	Principal	
-- Verify Placement	Fall	MT	Fall Goal Rpt.		Principal	
-- Curric. Implemt.	F-S	LT		CMI/class profile	Teacher	
3-- Indiv. Instruction	Fall	ST	Stdt. Goal Rpt.		Teacher	- needs goal level information by individual class
-- Assess program effectiveness	Spring	LT	Stdt. Goal Rpt.		Principal	- to download goal rpt. for RML
-- Reporting to Parents	F-S	MT	Stdt. Goal Rpt.		Teacher	- more data at individual student skill level
-- Problem w/ Stdt.	F-S	ST	Prin. Alpha/Rank		Admin Asst	level and for H-M-L goal data by class
-- Forecasting	Spring	LT	Rank Order Rpt		Principal	
-- Placement	Fall	MT		Locator test->RIT	Principal	
-- Check Progress	Fall	MT		Markham	Principal	
-- Evaluate by grade	Spring	LT	F-S Match Rpt.		Principal	
4-- Forecasting	Spring	LT	PALT data	tchr. observ.	Team lead.	- PPt type test for science
-- Placement	Fall	MT		Port.Prognostic	Test Coord	- school level data to review
-- Perform. Indicators	F-S	LT	Achiev.Profiles	district averages	Principal	overall performance
5-- Instruc. planning	F-S	MT	Stdt. Goal Rpt.		IS w/tchrs	- RITs are hard for parents;
-- Match mater.to goals	F-S	MT	Stdt. Goal Rpt.		Teachers	P-score is easier
-- Self-eval. Tching	F-S	LT	Stdt. Goal Rpt.		Some tchrs	
-- Indiv. Parent Conf.	F-S	ST	Test Score Cards		Principal	- TS Card is redundant to P/T Conf. Sheet
-- Indiv. Stdt. Perform	F-S	ST	Principals Alpha		Principal	
-- Answer Parent Quest.	F-S	ST	Principals Rank		Principal	
-- Teacher Planning	Spring	LT	F-S Match Rpt		Prin/Tchrs	
-- Class scheduling	Spring	MT	Feeder Rpt (alpha)		Principal	- arrange feeder in rank order
-- Set goals/emphasis and staff developmt	Fall	LT	Admin. Rpt #2		DOI w/Prin	
-- Obj. view of school	F-S	ST	Goal Progress Rpt		IS w/Prin	
6-- Set general goals	May-Jun	LT	Admin. Rpt #2		Principal	- format GGP by instructional group
-- Instruct. grouping	Spring	LT	GGP		Prin->Tchr	- TS Cards too timeconsuming;
-- Handicap placement	F-S	MT	Stdt. Goal Rpt.		Principal	arrange by instruc. group
-- Answer parent quest.	F-S	ST	Achiev. Profiles		Principal	- wants data, not graphs; they'll do graphs
-- Instruc. Regrouping	Spring	MT	GGP		Teacher	- student longitudinal rpt. for parent conf.
-- Instru. Team Plans	Fall	LT	Princ.Rank Rpt		Team	
-- Check goal progress	F-S	ST	Stdt. Goal Rpt.		Teacher	
-- Ind.Tchr. goal plans	Fall	MT	Stdt. Goal Rpt/GGP		Prin/Tchr	
-- Indiv. Stdt. Perform	F-S	MT	F-S Match Rpt		Prin/Tchr	

DECISION MAKING STRUCTURE

	DECISION TYPE	TIME	TERM	REPORT TYPE	OTHER DATA	WHO	NEEDS/CONCERNS
7--	TAG/Chi Placements	Fall	MT	Prelim. Prin. Rpt		Principal	- indiv. stdt. longitudinal report
--	Stdt. monitoring	Fall	MT	Stdt. Goal Rpt.		Prin/Tchr	
--	Check stdt progress	Spring	LT	F-S Match		Principal	
--	Weaknesses by grade	F-S	LT	Achiev.Profiles		Principal	
--	Instruc. Planning	Spring	MT		Indiv. test scores	Prin/Tchr	
--	Analyze goals	Apr-May	LT	GGP Graphic		Principal	
--	Instruc. Grouping	Spring	MT	Goal reports		Principal	
8--	Instru. Group Plans	Summer	LT	Goal Rpt.		Principal	- class goal rpt. useless for planning unless
--	Class/FTE plans	Summer	LT	Princ. rank order		Principal	sorted by instructional groups
--	Placement	Fall	MT	CAT		Principal	- need to get data sorted by course number
9--	Ind/Class Perform.	Spring	LT	GGP	PC download data	Principal	- wrong size for tchrs; they need clear,
--	Performance Trends	F-S	LT		RITs	Principal	sequenced and applicable information
--	Screening (for.lang.)	Spring	MT		RITs as benchmark	Principal	
--	Placements/Sched.	Fall	MT	Prelim. Prin. alpha		Principal	- needs 3 copies: Counselor/Adm.Asst./Office
--	Forecasting	Spring	LT	Feeder (rank)	tchr. judgemt.	Principal	
--	Graphics information	Fall	LT	Admin. Rpt #2		Principal	- data proliferation=confidentiality problem
10-	Instru. Grouping	Spring	LT	F-S Match	tchr. data	Prin/Tchr	- Test Score Cards not used at all
	" "			Stdt. Goal Rpt.		Prin/Tchr.	- high data use by Ch.1; other tchrs low
--	Planning review	Summer	MT	GGP and Admin Rpt #2		Principal	
--	Prin/Tchr Planning	Summer	LT	5 year graphs		Prin/Tchrs	
--	Regrouping/Placemt	Fall	MT	Prelim. Prin. Goal		Principal	
	" "	Fall	MT	Stdt. Goal Rpt		Principal	
11-	Instru. Grouping	Summer	MT	Goal Report		Principal	- Test Score Cards not useful at elem. level
--	Parent Conferences	F-S	MT	Par/Tchr Conf. Rpt		Prinicpal	
--	Staff/Curric Devel.	F-S	LT			Principal	
--	Indiv. Stdt Gain	S-F	MT	F-S Match		Prin/Tchr	
--	Instruc. Adjustments	F-S	ST		T judgmt/P obser	Teacher	
--	Indiv Stdt Confer.	F-S	ST	Prin. Alpha		Prin/Tchrs	- need alpha order by classroom level;
--	Group deficiencies	F-S	LT	Prin. Goal Rpt		Teacher	rank is not needed
12-	Set goal by subj/gr.	Summer	LT	Stdt. Goal Rpt		Prin/Tchr	- need data provided on cross-grade grouping
--	Planning/goal setting	Summer	LT	F-S Match		Prin/Tchr	- Test Score Cards on helpful
--	Regrouping	Fall	Lt	Prelim Prin Rpt		Prin/Tchr	
--	Parent Conferencing	F-S	ST	Prin Final (fall)		Principal	
--	Longitud. Perform.	F-S	LT	Admin. Rpt #2,		Principal	- convenient, compact format
13-	Instruct. Analysis	Spring	LT	Stdt. Goal Rpt		DOI/IS	- Test Score Cards not used in cluster
--	Show growth	Summer	LT	School Profiles/Achievmt Monitoring		IS	
--	Cluster planning	F-S	LT	Ethnic Rpt		IS	
--	Lesson/mat. plans	F-S	ST	Stdt. Goal Rpt		IS w/tchrs	
--	Parent Conferencing	F-S	ST		Peach brochure	IS	-Parents don't understand Parent-Tchr Guide
14-	Prog. Plan/Grouping	Spring	LT	Prin. Rank Rpt.		Principal	- combine R & L data to check relationships
--	Intervention Check	Spring	MT	Prin. Rank Rpt.		Principal	- Test Score Cards not used
--	Indiv Tchr Planning	Spring	LT	Stdt. Goal Rpt		Teacher	
--	Parent Conferences	F-S	ST	Prin. Rank Rpt		Principal	
--	Parent Conferences	F-S	ST	Par/Tchr Conf. Rpt.		Tchr/Par	- add school mean + GST RIT for each subj.
--	Summative Evaluation	Fall	LT	F-S Match Report		Prin/Tchrs	- would be better used if avail. in Spring
--	Placement/Planning	Spring	MT	Feeder Rpt.		Prin/Tchr	- develop a gain potential rating for stdts based on prior test scores to help placemt
--	Performance Summary	Fall	LT	Admin Rpt. # 2		DOI w/Prin	
--	Prog. Evaluation	F-S	MT	GGP		Prin/Staff	
--	Accountability Issue	F-S	LT		RITs->grades ques.	Prin/Tchr	- need stdt score + school and dist. means

PORTLAND PUBLIC SCHOOLS
501 North Dixon Street / Portland, Oregon 97227
Phone: (503) 249-2000
Mailing Address: P.O. Box 3107 / 97208-3107
OFFICE OF THE SUPERINTENDENT

Matthew W. Prophet
Superintendent

April 21, 1986

Dr. Verne A. Duncan
State Superintendent of Public Instruction
Oregon Department of Education
700 Pringle Parkway S.E.
Salem, OR 97310-0290

Dear Verne:

My staff and I, as requested, have recently reviewed and returned to your office corrected drafts of the Oregon School Profiles for the Portland schools. The changes required were substantial and we have appreciated the receptiveness of your staff to our recommendations for improvements in the accuracy and appropriateness of these sensitive documents. Our overall review of the state profiles, nevertheless, confirms the concerns about them that we and others throughout the state have been raising since they were first proposed. I will describe these concerns again, giving the reasons for them and drawing implications for change in future statewide school profiling efforts. Moreover, I will give the rationale for a strong request that the statewide assessment test data from the unrepresentative Portland sample of schools not be included in your current state profile record.

First, I want to say that I believe with you that local school profiling for local school improvement is a worthwhile educational practice. The issue here is how the state can best encourage and help local school districts to adopt and carry out such desirable practices.

As you know, the Portland School District has published its own school profiles for a number of years. As stated in your department's "Guidelines for the Development and Use of School Profiles," "the three major functions of the profile are to: 1) provide a focus for local school improvement efforts, 2) provide educators with data that stimulates (or assists with) program troubleshooting, and 3) assess the impact of school improvement efforts by examining the consequences of changes in policy, practices, and resources." In the years during which we have been profiling our schools, we have found that such profiles, if produced with accuracy and respect for legitimate variability among schools and if used with appropriate care, can indeed support the attainment of each of these functions, and thus serve very useful purposes. I therefore support you in your desire to use the resources of your office to make such profiling available for local school improvement to every district in Oregon.

There are, however, two major problems with the state profiling effort as currently conceived and executed:

1. It unnecessarily supplants and undermines in the name of statewide uniformity similar local profiling efforts which have been carefully thought out, developed, and refined over many years.

   As with many of the elements of the State Action Plan (such as curriculum aligned testing), Portland and other Oregon districts have earlier sensed the same need and have moved ahead assertively and creatively to meet it. Simply allowing local districts to add their own profiles will not solve the problem since the state profile addresses the same variables but often with sufficient difference in definition as to render the state and local data apparently contradictory. We understand and support the need for a strong state role in defining standards of quality education and educational management practices for local districts to meet and in providing leadership and technical support to those districts who have difficulty in responding. We believe, that the Oregon Department of Education can and should play a significant role in supporting local school districts in profiling schools as a step toward school improvement. For example, we find it more beneficial for our improvement planning to report school staff in counts of full- and part-time employees, whereas the State Profiles reports school F.T.E. While this distinction is relatively minor, the differences in the two counts can be confusing and counterproductive.

2. Because of the diversity of the schools and districts encompassed by the state profiles, the data offered for comparison is not and probably cannot be truly comparable and will thus lead to inappropriate and possibly destructive comparisons.

   For the sake of uniformity and efficiency, the state profiling system reports the same variables for each school, even when the data measuring those variables may be calculated or defined differently and may have widely ranging interpretations from district to district. For example, the calculation of the ratio of pupils to instructional staff means something entirely different in a district like ours that has a staffing model including pre-kindergarten instruction and classroom aides for early childhood enrichment, as compared to districts where such program support is not offered. There are strong and valid reasons for such differences among local districts but these are ignored and obscured by much of the data presented in the state profiles. Once the state-wide profiles become public documents, there

will be very little which can be done to limit misuse and abuse due to such noncomparability and other factors. We would wish to avoid any situations where the public might compare and evaluate schools based on a comparison of the data reported in the State Profiles. It would thus be better if no such single state profile were to be produced, but rather that all districts be required by standard to produce them according to clear but broad and flexible state guidelines and with state assistance as needed. If the currently underway statewide profiling system is to be implemented, my only suggested remedy with the inevitable problems of a uniform but non-comparable statewide profile is to proceed with extreme caution and to explain to all potential users the problems associated with them and the proper and improper uses of the current profile data. I believe this latter approach will not mitigate misuse, but give at least some rationale to counter the abuses when they happen.

I have an important specific request for the handling of one critical segment of data which was included in the draft of the state profiles which we edited. That draft contained spring 1985 Oregon State Assessment student test scores for the few Portland schools and grade levels selected for inclusion in the state sample. At the time that the State Assessment was conducted we pointed out that your sample of our schools was non-representative since it over-sampled predominantly low achieving schools. Your staff acknowledged the validity of this concern. Inclusion of Oregon test score data in the profiles (for which it was never intended) for only those non-representative schools and grades selected will unfairly and inappropriately stigmatize our district and our schools. Moreover, it could seriously impare the credibility our own well established test reporting and school profiling efforts. I ask that you take the steps necessary to ensure that the sample Portland school scores on the Oregon tests not become part of the state profile record.

I want to note again that your staff have been extremely cooperative and professional in developing the profile system and reporting the data. They have been open to suggestions for improvements and have made corrections when errors were pointed out. The problems that remain are not technical but are fundamental to the effort as conceived. Only you can remedy these problems by redirecting the effort and creating a state-local partnership approach using standard setting, accountability, local flexibility, and technical assistance. You can thus take a firm, aggressive leadership position and give direction that will move schools and children toward excellence without feeding the media's desire for controversy or the public mythology about weakening school productivity.

Again, I want to express my appreciation to you and your staff for your responsiveness in dealing with our concerns with this and other key elements of The Action Plan. It is good that we have been able to work together up to this point to avoid critical difficulties. Continuation of our collaborative approach can guarantee that both state and local needs are excellently met within the framework of the strong state plan.

Sincerely,

Matthew W. Prophet
Superintendent

MWP/wp

APPENDIX V

A REGIONAL AND LOCAL ITEM RESPONSE THEORY BASED TEST ITEM BANK SYSTEM

Walter Hathaway, Ron Houser, and Gage Kingsbury

Portland Public Schools
Department of Research and Evaluation
Portland, Oregon

November 1985

INTRODUCTION

A decade and a half ago, a small group of researchers and test developers in the Portland, Oregon School District realized the potential of Item Response Theory for developing calibrated item banks that would be tools for better educational measurement. The purposes we wanted our citywide testing programs to support included equitable, effective, efficient, valid, and reliable:

- Grouping and placing students
- Targeting instruction on individual student learning needs
- Evaluating student progress over time
- Identifying neglected areas of the aligned curriculum and evaluating and improving programs and services at the student, classroom, grade within school, and grade within district levels
- Providing accountability to the school board and the community

In order to meet these needs, we needed an educational measurement system that would answer the following questions:

- Is the current rate of gain of this student, class, grade, or program satisfactory compared to his/her/its age, grade, program mates and the previous pattern of gains observed?

- What are the current strengths and weaknesses (in terms of goal areas needing further diagnoses and possible work) of this student, class, grade, school, or program, and how have they changed over time?

- Is the level at which this student, class, grade, or program currently performing satisfactory compared to his/her/its age, grade, or program mates and the previous pattern of levels observed?

We could not find any available measurement program that would help us answer these questions and meet these needs adequately, and so we set out to build one ourselves.

There followed a period of extensive collaborative research and development, much of it within the framework of the Northwest Evaluation Association, which was created to foster regional cooperation in and mutual benefit from this effort. The result today is a system of three comprehensive basic skills item banks in Reading, Mathematics, and Language Usage. The constantly growing item banks in Reading and Language Usage each have over 2,000 field tested, calibrated items linked to a common, continuous curriculum scale for each subject. The Mathematics item bank now has over 3,000 such items. State and local school systems have been using these item banks since 1977 to construct effective, efficient survey achievement tests, competency tests, and other instruments that combine the best qualities of criterion referenced and norm referenced measurement. These excellent measurement systems have been the cornerstones of state, district and school renewal efforts that anticipated "A Nation at Risk" (Gardner, 1983) by at least five years. The ongoing collaboration is now resulting in a similar item bank in Science, and yet

another in Social Studies is on the drawing board. Computerized adaptive versions of the tests based upon the item banks are now being pilot tested in Portland.

DESCRIPTION OF THE CURRENT ITEM BANKS

At present, there are approximately 2,200 items that have met all the pre-screening, editing and statistical criteria in Reading; 2,300 items in Language Usage; and 3,500 in Mathematics. Each year 1,800 new items are field tested across the three subject areas, with about 1,200 being added to the item banking systems. Items are added to the banks by a process called the "Developmental Testing Program." Within two weeks of sitting for a regular Achievement Level Test, all students in grades three through eight also sit for a short test of between 20 and 25 items in either Reading, Mathematics, or Language Usage.

In an attempt to more reliably calibrate items on a short test with the current item banks, a procedure is employed that is termed the "fixed parameter model" (Houser, Hathaway & Ingebo, 1983). Since the students have a relatively reliable estimate of their ability through the Achievement Level Testing Program, their ability estimate can be essentially "fixed" for the Development Test at that point on the underlying metric. Item difficulty is the only parameter that is necessary to estimate. Thus, one can calibrate a one item test with the same precision as is achieved on a functional level test.

In cooperation with the Northwest Evaluation Association Science Project, Portland Public Schools is currently in the process of developing a Science item pool which now consists of about 3,500 items.* These items have not been scaled to a common metric at the present time; however, the project is currently engaged in the first phase of a research project that will help determine the number of scales needed and allow the project to begin the design of a series of field tests leading to a final set of scaled Science item banks. The project has plans to scale as many as 16,000 items by the year 1990.

WHY THE ITEM BANKS WERE DEVELOPED

Advantages of Item Response Theory

Classical test theory has been the backbone of educational and psychological testing for most of this century. The concepts of reliability and validity have enabled the development of tests of psychological traits based on sound theory and practical experience. In the last two decades though, testing techniques have been developed and refined to use new theoretical models to describe the interaction between the test-taker and the individual test questions.

---

*The term "pool" was used to distinguish the set of items which do not have a common metric. The term "bank" has been reserved for items which are scaled to a common metric.

These item response theory (IRT: Lord & Novick, 1968; Lord, 1980) models allow the estimation of differences in the measurement characteristics of test questions. These item characteristics may then be used to determine an individual's score on a test and the precision of the score.

The major advantage of these IRT models over classical test theory is that classical item and test characteristics vary, depending upon the group of students taking the test, while IRT item and test characteristics do not. Classical indices of item difficulty, point-biserial correlation, and test-retest reliability may all change when groups of test-takers differ in mean ability or ability distribution. When IRT models are used, item parameters are not biased by the specific sample of students taking the test or by the other items included on the test reference. In addition person parameters are not biased by the particular form of the test taken or by the other students taking the test.

This sample independence means that, except for differences in precision, the student samples used to estimate item characteristics are interchangeable, and the item samples used to estimate student characteristics are interchangeable. It should be noted, however, that these properties of IRT only pertain if the items and students are sampled in some manner from the populations of interest used to derive the IRT scale (Lord & Novick, 1968).

A second advantage of IRT over classical test theory is that a measure of the precision of measurement is available for any question or set of questions with known item characteristics. This standard error differs from that found in classical test theory in that it varies as a function of the student's

achievement level. This is a much more rational result than the constant standard error obtained in classical test theory since we know that the error is not the same for individuals scoring, e.g., 10, 25, and 40 on a 50-item test.

The result of these properties of sample independence and a known functional form for the standard error is that test developers can develop test forms with confidence in the comparability of scores and with some knowledge of the precision of the scores that will be observed. Any set of items taken from a bank of items calibrated using this procedure will produce a scale score and correspondence error of measurement. The size of the error depends on which items are selected, how many items are selected, and the student's raw score on that set of items.

Item Banks and IRT: A Common Scale

A second consequence of the properties of the IRT models is that large groups of items addressing the same subject can be brought together onto a common scale. This makes the IRT models invaluable to anyone trying to create or maintain a large item bank. Large item collections may be easily brought together for use with classical test theory (as in domain-referenced testing); but in order to create anything but randomly parallel test forms, some type of overall scale is a necessity.

The linking procedures and designs which are a direct outgrowth of IRT theory allow researchers to scale thousands of items together (Hathaway, 1980).

From these item banks tests for different purposes and measurement of students' different functional levels may be drawn. Without this common scale the usefulness of a large collection of test questions is severely limited.

THE DEVELOPMENT AND EXPANSION OF THE ITEM BANKS

In the early 1970's, the Portland School District's Research and Evaluation Department, inspired by the work of Georg Rasch and Benjamin Wright, began investigating the applicability of IRT models to the development of item banks from which achievement tests in Reading, Mathematics, and Language Usage could be produced. The Rasch model or one-parameter IRT model was identified as the most promising of these models at the time because of its more advanced stage of development as well as its conceptual clarity and practicality. Later in the decade, while other test developers moved to two- and three-parameter models, Portland's Research and Evaluation Department concentrated its efforts on improving methods of linking items in Rasch-calibrated tests to a single, continuous scale; on reducing the need for second-parameter correction by discarding items that did not fit the model; and on reducing the need for third-parameter correction by readministering higher or lower level tests to students not performing at "percent right" levels that fell within the measuring range of the tests.

Before accepting the Rasch model as a useful tool for the construction of item banks and tests, the Portland Research and Evaluation Department conducted extensive research on such questions as:

1. Does the calibration procedure yield the same scale values regardless of the student sample?

2. How many students are needed to scale test items within acceptable error limits?

3. Does the calibration procedure used produce comparable achievement levels for the same students on different tests?

4. Does an item receive the same scaling regardless of other items in the test?

Once the initial basic research had been completed, work began in earnest on collecting, writing, screening, cataloguing, field testing, calibrating, statistically analyzing, and linking items together into true item banks. Impetus was given to planning the content coverage of the banks by an earlier regional project called the Tri-County Course Development Project which developed, classified, and catalogued comprehensive sets of K-12 learning outcome statements in 14 subject matter areas including those covered by the first three item banks. Collaboration by state and local school districts in Oregon and Washington was gained within the framework of the Northwest Evaluation Association which proved a rich source of ideas, items, field test sites, and other critical, major contributions.

Item Sources

There are two procedures that are employed to identify new items to be field tested. First, items are identified from existing sources such as the National

Assessment of Educational Progress released items, items developed by other federally supported projects, and items shared by other state and local school systems. These sources were an extremely valuable and relatively inexpensive way to get potential new items in the early stages of the item bank development. Prior to the selection of items for each Developmental Testing Program, goal areas which are either lacking depth or some ranges of difficulty are identified. Over time, these "weak" areas have become fewer and thus the specifications for new items have become more restrictive. As this has occurred the public domain item sources have not been as productive a resource as they were in the initial phases of bank development.

Second, items are written to the specifications by trained item writers who usually are teachers. These teachers normally have considerable experience in teaching the level of students for whom they are to write new items. During the item writing process, teachers have the opportunity to critique one another's items. This process helps refine the end product, particularly when writing to a higher level of thinking goals. Considerable care is taken to help the item writers focus on district course goals rather than specific behavioral objectives. It is a policy of the Research and Evaluation Department that specific teaching methodologies not be measured. Rather, the more comprehensive learning goals to which those methodologies are directed are the guiding organization behind the development of the district-wide measurement program. The particular teaching strategies that should be employed in the classroom to help students meet the system's educational goals are left to the professional judgment of the individual teacher.

A number of format specifications are applied to screened and newly written items to ensure good item writing practices and to meet the need to have consistency among the items in a given bank. The sorts of things these guidelines address include: easily understandable directions, clarity, item responses of relatively the same length, good distractors, only one correct answer. If these specifications are not met, the items are either returned to the item writers or item screeners or, if the problem is a minor one, is simply modified by the item bank technical staff.

Items which meet the screening specifications are then assigned an identification number and are sent to the typist. Once the items have been initially formatted on a word processor, they are checked by a diverse team of professionals to determine whether they have any disqualifying gender or ethnic problems identified from an extensive checklist. Problem items are discarded or revised by the team and returned for correction. A final edit is made after all other corrections for any remaining format, spelling, grammar, and/or punctuation errors. The items are then grouped with other items assigned to the same grade level by the item writers or screeners.

Groups of items are then formatted into developmental field tests of 20 to 25 items and administered to students. Once the students' item responses are received, they are matched by student identification numbers to their Achievement Level Test score for that subject. A series of checks are made to determine if the match is correct and if the students did indeed respond to the form of the test indicated on their answer sheets. Upon satisfactory completion of these tests, the "fixed parameter" calibration analysis is run.

Item records are developed for each item on the test containing: the item number on the test, the item bank number, the keyed response, the test form number, the number of omits, the number of students responding to the item, the test date, the Rasch calibration, the calibration standard error, the point biserial correlation, the mean square fit, and the percent of students answering the item correctly. Item characteristic curves are also developed and printed for later evaluation.

The first item screening criterion used following testing is based on the number of students responding to the item. If this number is less than 200, the item is flagged for retesting in the next Developmental Program. The next criteria concerns the percent of students correctly answering the item. If this is less than 5 percent above chance level or greater than 90 percent, the item is flagged for retesting at either a higher or lower grade level as appropriate. In either of these two extremes it is felt that not enough of the item characteristic curve (ICC) is available to reliably determine its calibration. Next, the mean square fit (MSF) is checked, and if there is not a statistically significant misfit, it is then passed to the final statistical check. The last check is a visual inspection of the ICC to identify any abnormal deviations from the ideal model that was not picked up by the MSF.

Any item which fails the first two tests is automatically scheduled for retesting. Any item which fails the last two tests is sent back to a team of item writers for revision if the problem can be identified. If no obvious problem is identified, the item is deleted.

HOW WE CURRENTLY USE THE ITEM BANKS

Test Development

The Reading, Mathematics, and Language Usage item banks are the prime resource employed in the development of the various functional levels tests administered to all students in the Portland Public Schools, grades three through eight, in both Fall and Spring (see Haney, 1985, for a review). Two series of tests in each subject area are designed to measure all six grade levels except in Mathematics, where the instructional sequence of the curriculum dictated the development of two series for each Fall and Spring program so that high achieving younger students would not be expected to respond to items measuring goals for which they have had no instruction.

Each series consists of multiple-level tests of different difficulty. Each level is 20 RITS (2 logits) wide in difficulty. (For students in a particular grade, the standard deviation of achievement levels is about 15 RIT points.) A level's content overlaps the previous and next higher level by 50 percent. A minimum of seven items is necessary to report a subscore and a total test consists of between 45 and 60 items. Students are placed into these levels by either a short locator test, teacher judgment, or by any or all of their previous six scores.

The item banks also comprise the primary resource for the construction of the Graduation Standards Testing Program (Hathaway, 1980). These tests are

designed to make competency decisions and so are focused on a cutoff point on the underlying metric. Most of the items are centered around the cutoff point so that the best measurement is obtained at that point. However, an extension of the measurement downward is achieved by including items which are less difficult. This is done so that goal-based information is available to help identify weak areas for students not performing above the cutoff point.

Computerized Adaptive Testing

To allow more school-based control of testing and to provide school personnel with a capability to use the techniques of modern test theory, we have been investigating the uses of computerized adaptive testing (CAT: Weiss & Kingsbury, 1984).

CAT is a technique for testing which employs a computer to present the most appropriate test questions to a student. The student answers questions on the computer keyboard and the computer scores and records the student's responses. After each item the computer searches an entire item pool (previously stored in the computer) and selects the next test question for the student. At each step in the test, the computer refines its estimate of the student's achievement level and gives the student the question left in the item pool that will provide the maximum information.

From the student's point of view, the test questions become more difficult as questions are answered correctly and less difficult as questions are answered incorrectly. This is akin to the levels assignments done in the functional

levels tests described above, but in a more direct manner, based on the student's current performance.

The CAT system will help serve the schools' special needs for interim measurement, initial pupil placement, and individual graduation standards testing. It is hoped that this form of testing will allow teachers and administrators to assess gains from individualized instruction that occur between the district-wide assessments that occur in the Fall and the Spring. It is expected that the CAT system will serve as a useful addition to the current testing program and help us come one step closer to a continuous measurement model.

The present pilot project, active in six schools ranging from elementary to high schools, is designed to assess the reactions of teachers and students to CAT, and to provide assurance that the CAT system measures the basic skills in the same manner as the current, district-wide, paper-and-pencil tests. Future research will address the potential for enhanced measurement with the CAT system.

PROBLEMS AND ISSUES IN ITEM BANK DEVELOPMENT, MANAGEMENT, AND USE

Choice of an IRT Model

Although the one-parameter model is currently being used with the basic skills item banks, we have not rejected the possibility of the use of other item

response models in the future. As mentioned above, the one-parameter model was originally chosen for use because of its strong theoretical and practical development. If another model were found or developed to the point at which it was demonstrated to have substantial, practical advantage over the one-parameter model, it could be used to recalibrate the current item banks or more likely be applied to the item banks currently under development.

Aside from the one-parameter model based on item difficulty, researchers have developed a two-parameter model which adds an item parameter allowing differences in item discriminatory power and a three-parameter model which adds an item parameter allowing differences in the lowest expected probability of a correct response.

These models differ in three respects. These are the estimation procedures used for item calibration, the scoring procedures used for estimating student achievement levels, and the extent to which student test responses conform to the model.

When building an item bank, it is essential that the response model chosen have practical procedures for the estimation of item and student characteristics. At the same time, the model should be flexible enough to capture the differences inherent in students and items. The one-parameter model has the edge in simplicity of estimation, particularly for student achievement levels. However, the more complex models may result in a more precise representation of item and student performance. Each item bank developer must weigh the advantages of each model with respect to the particular testing situation to determine the best IRT model.

When the basic skills items were first developed, a primary consideration was the number of students that would be needed to calibrate a given set of test questions. It became readily apparent that the only IRT model that would allow stable calibration and linking of items, and not overly restrict the growth of the item pools, would be the one-parameter model. This model allows stable and consistent estimation of item parameters with samples of 200 students. The two- and three-parameter models have required much larger sample sizes to obtain stable calibrations.

Recently, though, Swaminathan and Gifford (1985) have developed new techniques for estimation in the two-parameter model which may reduce the number of students needed to adequately estimate item parameters. In addition Lord (1983) has theoretically demonstrated improved measurement precision using the two-parameter model for all sample sizes greater than 200.

These studies as well as other research findings have led us to begin developing research designed to compare the practical impact of model choice on basic skills measurement.

Dynamism

To insure that the item banks continue to provide measurement reflective of the curriculum as it is being taught at any given time, the banks must be periodically checked and updated to correspond to these changes. There is some concern about the comparability of scores across time with the use of such a dynamic scale. However, metric comparability can be achieved by

employing those subsets of items which have calibrations that remain stable over time. The total measure would still be reflective of the underlying trait at any given point in time.

By adding new items to the item banks after every testing program and by including at least two testing programs in the annual calibration check, the entire item bank would slowly reflect major curriculum shifts. Calibration drifts over the years have generally been modest. The second district-wide administration of the Language Usage test did result in substantial calibration drift, however. To explain this finding, the hypothesis has been advanced that teachers were, for perhaps the first time, focusing instruction on the Language Usage goals.

Interaction with Curriculum

We were fortunate in Portland and in the Northwest in general to have a tradition of a commitment to goal-based instruction and measurement that extended back to the Sixties. We also had a history of strong curriculum and evaluation leadership and collaboration.

One of the greatest challenges, however, in developing the item banks and using them to create good, useful, curriculum aligned, instructionally sensitive measurement instruments has been to gain and maintain effective cooperation between Curriculum and Evaluation. Some of the principles and practices that have helped us succeed in this crucial but delicate part of the work are the following:

. We established and followed the principle that Curriculum would have the leadership and the final say so in determining what was to be measured, while Evaluation would lead on issues of how to measure.

. We shared control and resources. For instance, in Portland we formed a Curriculum and Evaluation Council made up of the senior personnel from each area. This group, which meets monthly, is chaired in alternate years by the heads of the Curriculum and Evaluation Departments. In the meetings information is shared, policy direction is agreed upon, resources are allocated, plans are made and monitored, task forces are created for such work as test blueprint design and reports, revisions, and recommendations of the task forces are reviewed, revised, and ratified.

. We learned how to negotiate (and even enjoy) good, clean conflict since our ends were always the same (namely more informed and better decisions about students, programs, and policies), whereas our disputes were about means which themselves could usually be put to the test of data through experimentation and pilot testing.

. We took the long view. Blessed by enlightened top administrators and school boards, we were able to count on the resources and support to plan and develop on five, ten, and even twenty-year horizons.

Quality Control

Quality control measures are employed to insure that all aspects of the testing program are functioning accurately and effectively. Objectives of quality control and means taken to accomplish them are described below:

1. Insure that schools receive pre-printed answer sheets that are organized as requested by each school (by homeroom, by subject matter, by class, by grade level, etc.) for test administration.

   The school master file showing how each school requested its answer sheets to be organized in the next prior testing is compared with the current set of request forms received from principals. Changes in these requests are identified (usually only a few) and the master file is updated. This file automatically governs the way pre-printed answer sheets are ordered for distribution to the schools.

2. Insure that booklets that are distributed are printed properly, have all pages, etc.

   Samples from each printing order are taken to insure that all pages are present, all items appear in the test, and quality of printing is satisfactory.

3. Insure that the quality of printing on the pre-printed answer sheets will permit accurate scanning.

Pre-printed answer sheets for each test administration are "fan scanned" to find any answer sheets in which the quality of print may not permit accurate scanning, and that all needed information is in its proper place.

4. Insure that the computer program for pre-printing name, I.D., grade level, and test code on answer sheets is functioning accurately.

All programs for pre-printing answer sheets are test-run and results are examined to insure that data are accurate and complete.

5. Insure that students are assigned the proper level of a test series.

The program that assigns students to their proper level tests in any given administration is test-run, and the results for a sample of students are examined in relation to previous test results to insure that the program is functioning properly. (Placement is based on an algorithm averaging each student's five prior test standard scores with double weighting on the last standard score. This average standard score is converted to a RIT score and level placement for the grade level and time of year the testing is being carried out.)

6. Insure that the students understand directions for taking the test.

Verbatim instructions for administering the PALT are provided every teacher, and the testing procedures involve completion of sample items to assure student understanding of how to use the machine-scored answer sheets. There are specially designed answer sheets and practice tests for use with beginning third grade students.

7. Insure that conditions of test taking are good.

Meetings are held with test building coordinators prior to each administration and guidelines for conditions of test taking, as outlined in manuals of instruction for coordinators and school personnel, are addressed at these meetings.

8. Ascertain validity of student performance.

Both the coordinator and teacher manuals state criteria for labelling answer sheets as invalid where student performance does not meet these criteria. In addition computer programs label as invalid any performance falling outside the valid measuring range of the test form the student is assigned to. Also, if a student has 10 or fewer item responses (items attempted) on an answer sheet, the computer notes that performance as invalid.

9. Insure that answer sheets are properly marked and completed.

Each teacher must complete a signed checkoff that he/she has inspected each answer sheet for erasures, multiple marks, extraneous marks, and use of inappropriate markers on the header sheet before the sheets are returned. Scanner programs are used when tests are scored to identify incomplete information for essential hand-coded data.

10. Insure that coded information on each answer is correct and complete.

The procedure described under 3 above insures that pre-coded information is correct.

11. Insure that students have taken the assigned form of the test.

For all students scoring below chance levels, each test is rescored using the scoring keys for all forms in the series. This identifies any student whose test form is different from the one coded on the answer sheet. In such cases the legitimate score is entered on the student's record.

12. Insure that students needing to be retested are retested (high and low).

Students scoring outside the valid measurement range of the test are made known to the school and assigned a higher or lower test form as appropriate (two levels higher or lower, if possible). Schools are provided the materials required for retesting and all schools routinely schedule such retesting following the regular program. Where retests are regarded as inappropriate by teachers, they are not required. In such cases low students receive no test record and high students retain the score achieved.

Complexity and Credibility

Making a complex system credible to the psychometrically unsophisticated is a major issue we have faced in developing IRT-based item banks and functional level testing systems based upon them. This becomes an especially critical

matter when the test results seem to suggest something negative about the performance of students, programs, etc. Rhetoric such as "voodoo testing" is only a breath away when low levels or small gains are registered by complex measurement systems.

There is no magic answer to this problem. Its solution must be approached on a variety of fronts, such as a good inservice program; frequent helpful communication with representatives of the media and with community groups; effective collaboration with public information departments; as well as production and dissemination of a layered system of media productions to tell the measurement system's story--ranging from brochures through workshops and manuals to effective use of electronic media such as video cassettes and cable television. It also helps to periodically have the system evaluated by independent measurement experts who can both sanction the program and suggest improvements.

WHERE WE ARE GOING

New Testing Directions

Since the development of the item bank capability, we have found ourselves continually being asked by teachers, principals, and curriculum specialists for more and better measurement instruments and test results reports. They are developing creative and productive ways to use good data on how students are performing relative to the goals of the district. It is indeed a refreshing change to have the requests for new information come from teachers and principals as well as from the School Board or Superintendent.

For instance, this year 42% of our Black students are reading above the national average up from 36% two years ago and 31% four years ago, an eleven-point gain. Since the national average is the point above which 50% of students achieve, we are now within eight percentage points of having this student population at the national level and are gaining at a rate at which that goal should be reached within three years.

We have been working to develop a scale for measuring attitudes toward the subject of mathematics, and have been exploring ways to create scales for other affective variables. Affective goals are a part of many school districts' curriculum continuums, and yet very little has been done to provide information back to teachers and other instructional leaders as to how effective they have been in developing those traits. We recognize the potential that measurement of affective traits could have in the classroom, but we have only started to develop one such scale and to explore the possibilities for others.

Currently, NWEA and many school districts including Washington, Oregon, California, and Missouri have begun a massive project of developing the capability of measuring Science goals in grades K through 12. Unlike the basic skills item banks, Science has a very diverse curriculum across the participating districts. This diversity produces constraints on the scaling process as well as pointing out a tremendous need to utilize the flexibility of item bank-based measurement by individual school districts.

We are currently examining data gathered in a field testing designed to explore the dimensionality of the underlying traits (Kingsbury, 1985) typically taught in Science. This data will provide the structure to begin

developing the scales for measurement. The five-year measurement objectives of the Science project are to develop as many as 16,000 items distributed across the Science goals. A particularly unique emphasis of the project is to focus not only on content-related items, but also on process and concept goals. Many concepts in Science transcend the particular content in which it is being taught (i.e., conservation, equilibrium cycle, etc.). Developing items to measure these concepts that are independent of the context of a particular subject area is meticulous and thus time consuming.

New Testing Technology

In an effort to allow teachers and school administrators to take advantage of the item banks and modern testing procedures, further applications of computers in the classroom are being explored. As an example, systems are currently being designed to allow the storage and retrieval of test questions from a video disk.

Since the storage capacity of a video disk is so great, this would allow the efficient use of very large item banks. In addition the capability of the video disk to incorporate motion and sound will allow the expansion of the current item banks to include simulation questions, progressive response questions, and questions which do not force the student to read. This will allow the expansion of the CAT program to the more accurate measurement of the basic skills and the expansion of our current measurement to higher level skills.

Another direction of development with the item banks is the technology to allow computerized test design and generation. This new capacity will allow district-wide tests to be developed and produced in a more efficient manner. In addition this technology could be used in the future by teachers and administrators to directly access the item banks to enhance school-based decision making. This use of the computer to store and maintain item content and statistics will make the problem of administering item banks somewhat more manageable.

A basic goal of the item bank work is to close the gap between testing and instruction. One phase of this is to produce tests which have more direct instructional consequence. The CAT pilot project is a first attempt in this direction. By developing tests which more efficiently test the basic skills, more testing time is left to perform pre-diagnostic and diagnostic testing. The ultimate goal of this research effort is to create a more viable interface between the teacher and the evaluator (in this case in the form of the computer).

HOW OTHERS CAN SHARE

Readers who have come this far may be wondering how they can gain access to the item banks, tests, and technology we have been describing and, perhaps, how they can participate in their further development. Here are some answers:

- The current Reading, Mathematics, and Language Usage banks are being disseminated to local school districts for their own use by the Northwest Evaluation Association under a license from the Portland School District. Contact can be made by writing:

Ray Miller, Executive Secretary
NWEA
ESD #121
1410 South 200th
Seattle, WA 98148

or by phoning: 206-242-9400, Ext. 58, or 206-839-3932.

Participation in the ongoing collaboration within the NWEA framework can be initiated through the same source.

- Licenses for use of the current three basic skills item banks by others than local school districts may be requested from the Portland School District by contacting:

Dr. Walter E. Hathaway, Director
Research and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 206

- Information concerning participation in the NWEA Science Curriculum and Assessment project may be obtained by contacting:

Susan Smoyer
NWEA Science Project
700 Pringle Parkway S.E.
Salem, OR 97310
(503) 378-4157

- Questions concerning the Computerized Adaptive Testing project and other applications of modern technology to testing should be addressed to:

Dr. G. Gage Kingsbury
Research and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 229

- Information about the Portland Achievement Levels Tests and the application of IRT to the item banks may be obtained from:

Dr. Ronald L. Houser
Research and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 253

- The Portland Achievement Levels Tests and the scoring and reporting software are published and marketed by Microprocessors for Education and Business, 109 N. Main Street, Gresham, OR 97030 (503-666-7883).

- The results of the past and ongoing research and development related to our efforts are available in the form of reports, monographs, conference presentations, articles in popular and refereed journals,

book chapters and books, as well as on electronic networks including the American Educational Research Association's COMPUSERV system and The Council of Great City Schools' Telenet.

We welcome the interest and involvement of our fellow researchers, developers, and educators in this important work.

References

Gardner, D. P. (Chair) (1983). A Nation at Risk: The Imperative for Educational Reform. Washington, D.C.: U.S. Government Printing Office.

Haney, W. (1985). Making Testing More Educational. Educational Leadership, October, 4-13.

Hathaway, W. E. (1980). A School-District-Developed, Rasch-Based Approach to Minimum Competency Achievement Testing. In R. M. Jaeger & C. K. Tittle (Eds.) Minimum Competency Achievement Testing: Motives, Models, Measures, and Consequences. Beverly Hills, CA: McCutchan.

Hathaway, W. E. (Ed.) (1983). Testing in the Schools: New Directions for Testing and Measurement, Number 19. San Francisco, CA: Jossey-Bass.

Houser, R. L., Hathaway, W. E., & Ingebo, G. S. (1983). An Alternate Procedure to Obtain Ability Estimates in Latent Trait Models. Paper presented to the annual meeting of the American Educational Research Association, Montreal, Canada.

Kingsbury, G. G. (1985). A Comparison of Item Response Theory Procedures for Assessing Response Dimensionality. Paper presented to the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Small N Justifies Rasch Model. In Weiss, D. J. (Ed.) New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. New York: Academic Press.

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Swaminathan, H. & Gifford, J. A. (1985). Bayesian estimation in the two-parameter model. Psychometrika, 50, 349-364.

Weiss, D. J. & Kingsbury, G. G. (1984). Applications of Computerized Adaptive Testing to Instructional Problems. Journal of Educational Measurement, 21, 361-375.

0004E
11/21/85

SEMINAL LITERATURE

REFERENCES

Wright, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. and Douglas, G.A. Better procedures for sample-free item analysis. Research Memorandum No. 20, Statistical Laboratory, Department of Education, University of Chicago, 1975b.

Wright, B.D. and Douglas, G.A. Rasch item analysis by hand. Research Memorandum No. 21, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B.D. and Douglas, G.A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977a, 1, 281-294.

Wright, B.D. and Douglas, G.A. Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement, 1977b, 37, 573-586.

Wright, B.D. and Mead, R.J. CALFIT: Sample-free calibration with a Rasch measurement model. Research Memorandum No. 18. Statistical Laboratory, Department of Education, University. of Chicago, 1975.

Wright, B.D. and Mead, R.J. BICAL: Calibrating items with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright B.D., Mead, R.J. and Draba, R.E. Detecting and correcting test item bias with a logistic response model. Research Memorandum No. 22, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B.D. and Panchapakesan, N. A procedure for sample-free item analysis, Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B.D. and Stone, M.H. Best Test Design, MESA Press, Chicago, 1979.